BAB I
PENDAHULUAN

1.1. Latar Belakang

Klasifikasi teks merupakan teknik yang digunakan untuk mengelola data digital dan merupakan bagian penting dalam pemrosesan bahasa alami. Metode ini memungkinkan pengelompokan data teks secara otomatis ke dalam kategori yang telah ditentukan sebelumnya, sehingga mengeliminasi kebutuhan untuk klasifikasi secara manual yang cenderung lebih mahal dan memerlukan waktu lebih lama. Implementasi klasifikasi teks dapat ditemukan dalam berbagai bidang, seperti pendeteksian berita bohong, pengambilan keputusan, ekstraksi informasi dari data mentah, dan berbagai aplikasi lainnya [1]. Seiring dengan meningkatnya volume dokumen teks yang dihasilkan setiap hari, kebutuhan akan metode klasifikasi teks juga semakin meningkat. Jumlah data yang terus bertambah jauh melampaui kapasitas manusia dalam mencari, mencerna, dan menggunakannya secara manual [2]. Oleh karena itu, berbagai inovasi dan optimalisasi terus dilakukan guna meningkatkan akurasi serta efisiensi dalam pemrosesan dan pengelolaan data teks secara otomatis.

Dalam perkembangannya, terdapat dua kategori utama dalam algoritma klasifikasi teks, yaitu model tradisional dan model pembelajaran dalam (*deep learning*). Model tradisional bergantung pada efektivitas ekstraksi manual dengan contoh algoritma seperti *Naive Bayes, Support Vector Machine*, dan *Random Forest* [1]. Sementara itu, model pembelajaran mampu mempelajari representasi fitur yang lebih kompleks secara otomatis, sehingga dapat mengurangi ketergantungan terhadap teknik ekstraksi fitur manual [1]. Contoh algoritma yang termasuk di dalamnya adalah *Convolutional Neural Network* dan *Bidirectional Encoder Representations from Transformers*. Selain dua kategori tersebut, terdapat pula algoritma *ensemble* yang mengombinasikan berbagai model untuk meningkatkan kinerja klasifikasi teks. Salah satu teknik *ensemble* yang banyak digunakan adalah *boosting*, yang bertujuan untuk meningkatkan akurasi dengan menggabungkan beberapa model lemah menjadi satu model yang lebih kuat [3]. Teknik ini dirancang untuk mengurangi bias dan meningkatkan akurasi, menjadikannya metode yang efektif dalam menangani data teks yang kompleks [3].

Twitter adalah media berbagi pesan singkat dan jejaring sosial waktu nyata (*real time*) yang banyak digunakan secara global dalam kurun waktu satu dekade terakhir [4]. Berdasarkan data dari Radio Republik Indonesia (RRI), Indonesia menempati peringkat keempat sebagai negara dengan pengguna Twitter terbanyak pada tahun 2024, dengan jumlah pengguna mencapai 24,85 juta orang [5]. Hal tersebut menunjukkan bahwa Twitter tidak hanya berfungsi sebagai sarana interaksi sosial, tetapi juga sebagai sumber informasi yang kaya, mengingat banyak pengguna memanfaatkannya untuk berbagi beragam konten, mulai dari opini, berita, hingga informasi terkini. Akibat tingginya aktivitas tersebut, setiap harinya dihasilkan lebih dari 500 juta *tweet* yang merepresentasikan berbagai topik dan peristiwa [4]. Selain itu, Twitter sering dijadikan objek penelitian karena karakteristiknya yang bersifat cepat, dengan pembatasan 280 karakter pada setiap unggahan, sehingga informasi dapat disampaikan secara singkat dan padat [6].

Twitter, dengan karakteristik penyampaian informasi yang cepat dan *real time*, sering dimanfaatkan sebagai media untuk menyebarkan informasi terkait peristiwa darurat, termasuk bencana alam. Kondisi ini menjadi semakin relevan mengingat Indonesia dikenal sebagai supermarket bencana karena memiliki tingkat kerawanan yang tinggi terhadap berbagai jenis bencana [7]. Informasi yang dibagikan umumnya mencakup lebih dari satu aspek dalam satu pesan. Misalnya sebuah *tweet* dapat sekaligus menyebutkan lokasi kejadian, jumlah korban, serta kebutuhan bantuan. Oleh karena itu, pendekatan klasifikasi multilabel dipandang lebih sesuai dibandingkan klasifikasi tunggal, karena mampu mengakomodasi kondisi di mana satu teks mengandung beberapa kategori informasi secara bersamaan. Dengan demikian, hasil klasifikasi menjadi lebih representatif terhadap kompleksitas informasi di media sosial, serta lebih bermanfaat untuk mendukung pengambilan keputusan dalam respons bencana.

Penelitian ini berfokus pada empat label utama, yaitu lokasi, kerusakan, korban, dan bantuan. Relevansi pemilihan label tersebut didukung oleh berbagai penelitian sebelumnya. Studi terdahulu menunjukkan bahwa informasi mengenai kerusakan infrastruktur dan korban manusia merupakan kategori dominan yang dapat diekstraksi dari unggahan media sosial saat bencana [8]. Penelitian lain menegaskan pentingnya identifikasi lokasi korban serta jenis bantuan yang

dibutuhkan, seperti evakuasi, makanan, atau layanan kesehatan, untuk mendukung efektivitas penanganan darurat [9]. Lebih lanjut, penelitian terkini juga menggarisbawahi bahwa klasifikasi unggahan bencana berdasarkan kategori korban dan kerusakan memberikan kontribusi signifikan dalam memahami dampak bencana secara waktu nyata [10]. Selain itu, terdapat juga label lainnya yang digunakan untuk data yang tidak mengandung informasi empat label tersebut.

Categorical Boosting (*CatBoost*) merupakan salah satu algoritma *boosting* yang memiliki kemampuan sangat baik dalam klasifikasi berbagai jenis data, termasuk data teks seperti komentar [11]. *CatBoost* bekerja dengan membangun dan menggabungkan beberapa algoritma *Decision Tree* untuk menghasilkan prediksi akhir. Keunggulan utama algoritma ini terletak pada kemampuannya mengurangi *overfitting* melalui mekanisme *ordered boosting* [12]. Kemampuan ini sangat relevan untuk data Twitter yang jumlahnya terbatas namun kompleks, sehingga model tetap mampu melakukan generalisasi dengan baik. Selain itu, *CatBoost* mmemiliki metode *ordered target statistics* untuk mengolah fitur kategorikal [12]. Dari sisi kualitas teks, *tweet* umumnya bersifat bising (*noisy*) serta menghasilkan representasi fitur berdimensi tinggi setelah tahap *preprocessing*. *CatBoost* mampu mengatasi kondisi ini karena sifat *boosting* yang iteratif memperbaiki kesalahan prediksi secara bertahap. Lebih lanjut, dengan dukungan *MultiOutputClassifier*, *CatBoost* dapat digunakan untuk klasifikasi multilabel dengan hasil yang stabil dan akurat.

Algoritma *CatBoost* memiliki sejumlah parameter penting yang dapat disesuaikan, seperti *learning rate*, *depth*, dan *iterations*, yang secara signifikan memengaruhi kinerja model. Optimasi *Bayesian* digunakan untuk memperoleh kombinasi parameter tersebut yang dapat meningkatkan akurasi dan efisiensi model. Keunggulan utama teknik ini adalah kemampuannya mengeksplorasi ruang parameter secara lebih efisien dibandingkan metode pencarian konvensional. Hal ini dimungkinkan karena optimasi *Bayesian* memanfaatkan proses *Gaussian* untuk memodelkan ketidakpastian dan memprediksi area yang paling potensial menghasilkan perbaikan . Keunggulan tersebut sangat relevan dalam konteks klasifikasi multilabel data Twitter, yang umumnya berdimensi tinggi dan memiliki

distribusi label tidak seimbang. Oleh karena itu, strategi pencarian parameter ini mampu meningkatkan akurasi sekaligus menjaga efisiensi komputasi.

Setelah proses optimasi dilakukan, diperlukan evaluasi kinerja model baik sebelum maupun sesudah penyetelan parameter. Pada penelitian ini, evaluasi dilakukan dengan menggunakan tiga metrik, yaitu *hamming loss*, *f1-score weighted*, dan *subset accuracy*. *F1-score weighted* digunakan untuk mengevaluasi performa model klasifikasi dengan mempertimbangkan keseimbangan antara *precision* dan *recall*, serta memperhitungkan distribusi jumlah label [13]. *Subset accuracy* merupakan metrik yang lebih ketat untuk kasus klasifikasi multilabel karena hanya memberikan nilai benar jika seluruh label pada suatu instance diprediksi secara tepat [13]. Sementara itu, *hamming loss* digunakan untuk mengukur rata-rata kesalahan prediksi pada setiap label secara individual, yang menjadi indikator penting dalam konteks klasifikasi multilabel [14].

Berdasarkan penelitian sebelumnya, klasifikasi multilabel pada teks telah diterapkan dalam berbagai topik, seperti hadis Bukhari dan dokumen artikel. Penelitian-penelitian tersebut memberikan wawasan yang berharga mengenai teknik klasifikasi multilabel pada data teks. Salah satu pendekatan yang digunakan adalah *Binary Relevance*, yang memungkinkan algoritma seperti *K-Nearest Neighbor* dan *AdaBoost* [2] untuk melakukan klasifikasi multilabel. Selain itu, tahap *stemming* perlu dipertimbangkan dengan baik, karena jika tidak sesuai dengan karakteristik data, dapat menghilangkan informasi penting dan justru menurunkan performa model [15]. Pendekatan lain untuk klasifikasi multilabel yang juga dapat digunakan adalah *Classifier Chain* dan *Multi-Output Classifier* [13]. Di sisi lain, penelitian sebelumnya telah menunjukkan bahwa *CatBoost* memiliki performa lebih baik dibandingkan algoritma *boosting* lainnya, karena struktur pohon simetris yang lebih cocok untuk data non-linear [12]. Optimasi *Bayesian* juga terbukti mampu meningkatkan performa *CatBoost* dengan menemukan parameter terbaik untuk model tersebut [11].

Penelitian ini berfokus pada pengembangan algoritma klasifikasi multilabel untuk data teks, khususnya komentar Twitter terkait bencana alam. Meskipun berbagai algoritma telah digunakan untuk klasifikasi multilabel, pemanfaatan *CatBoost* masih terbatas, padahal algoritma ini memiliki keunggulan dalam

menangani fitur kategorikal, mengurangi *overfitting* melalui struktur pohon simetris, serta efisien dalam memodelkan data bising (*noisy*). Keunggulan tersebut semakin optimal dengan penerapan *Ordered Target Statistics* (OTS) untuk konversi fitur kategorikal, *Ordered Boosting* untuk meminimalkan bias, serta optimasi *Bayesian* untuk memperoleh parameter terbaik. Selain itu, Twitter dipilih sebagai objek penelitian karena bersifat *real time*, banyak digunakan masyarakat Indonesia saat terjadi bencana, dan informasinya umumnya mencakup lebih dari satu aspek, sehingga pendekatan klasifikasi multilabel lebih sesuai dibandingkan klasifikasi tunggal.

Urgensi akademik penelitian ini terletak pada penyediaan referensi mengenai penerapan *CatBoost* dengan optimasi *Bayesian* untuk klasifikasi multilabel pada data Twitter terkait kebencanaan, yang hingga kini masih jarang dilakukan. Selain itu, penelitian ini juga menghasilkan aplikasi berbasis Streamlit sebagai *Graphical User Interface* (GUI) yang memudahkan pengguna dalam mengakses hasil klasifikasi. Melalui GUI ini, pengguna dapat mengetahui jenis informasi apa saja yang terkandung di dalam *tweet*, meskipun hanya sebatas klasifikasi pada label (verifikasi bencana, lokasi, kerusakan, korban, dan bantuan). Aplikasi ini juga dilengkapi dengan *dashboard* interaktif yang menyajikan gambaran umum data, termasuk distribusi label dan pola informasi, sehingga hasil penelitian tidak hanya sebatas membangun model, tetapi juga dapat divisualisasikan secara komprehensif.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dibuat rumusan masalah sebagai berikut.

1. Bagaimana pengumpulan data Twitter mengenai bencana alam dari tahun 2020-2024 untuk klasifikasi informasi *tweet* bencana alam?
2. Bagaimana melakukan *preprocessing tweet* bencana alam yang telah dikumpulkan sehingga siap digunakan untuk pemodelan?
3. Bagaimana proses pelabelan informasi *tweet* bencana alam?
4. Bagaimana melatih model *Categorical Boosting* dan optimasi *Bayesian* untuk klasifikasi multilabel informasi *tweet* bencana alam?

5. Bagaimana mengukur performa algoritma *Categorical Boosting* dan optimasi *Bayesian* untuk klasifikasi multilabel *tweet* bencana alam menggunakan *hamming loss* dan *confusion matrix*?
6. Bagaimana menerapkan hasil klasifikasi informasi *tweet* bencana alam dalam bentuk *GUI*?

1.3. Batasan Masalah

Batasan masalah yang terdapat dalam penelitian ini adalah sebagai berikut.

1. Penelitian ini menggunakan 4898 data komentar mengenai banjir dan tanah longsor.
2. Data yang diambil merupakan komentar bencana alam selama tahun 2020-2024.
3. Pelabelan data dilakukan ke dalam kategori bencana, lokasi, kerusakan, korban, bantuan, dan lainnya.
4. Metode yang digunakan adalah *Categorical Boosting* dan optimasi *Bayesian*.

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, penelitian ini bertujuan untuk melakukan klasifikasi multilabel informasi *tweet* bencana alam menggunakan *Categorical Boosting* dan Optimasi *Bayesian*.

1.5. Manfaat Penelitian

Penelitian mengenai identifikasi bantuan korban bencana alam ini diharapkan dapat memberikan manfaat kepada berbagai pihak.

1. Universitas

   Penelitian ini turut berkontribusi dalam pengembangan ilmu pengetahuan dan teknologi, terutama di bidang *data science* dan machine learning. Dengan menerapkan algoritma *CatBoost* untuk menangani permasalahan klasifikasi multilabel serta menggunakan optimasi *Bayesian* dalam proses pencarian *hyperparameter* optimal, penelitian ini memperkaya referensi akademik terkait penggunaan model-model canggih yang efektif dan akurat dalam memproses

data dengan banyak label. Selain memperdalam kajian di area *Natural Language Processing* (NLP), penelitian ini juga membuka peluang bagi civitas akademika untuk mengembangkan lebih lanjut algoritma berbasis *gradient boosting* dalam analisis data teks, khususnya yang bersumber dari media sosial seperti Twitter.

2. Masyarakat

   Penelitian ini menyediakan informasi kebencanaan yang lebih terstruktur dan tepat. Melalui penerapan klasifikasi multilabel pada data yang diperoleh dari media sosial, berbagai informasi penting seperti jumlah korban, tingkat kerusakan, serta kebutuhan bantuan dapat dikelompokkan dengan lebih sistematis. Meskipun sistem ini belum beroperasi secara *real time*, hasil klasifikasi tetap memberikan gambaran yang lebih jelas mengenai dampak suatu bencana. Pada masa mendatang, sistem ini memiliki peluang untuk dikembangkan menjadi platform yang mampu menyajikan informasi secara langsung, sehingga dapat meningkatkan kesiapsiagaan serta respons masyarakat terhadap bencana. Selain itu, penelitian ini juga berkontribusi dalam meningkatkan pemahaman masyarakat terhadap pemanfaatan media sosial sebagai sumber data yang bermanfaat untuk tujuan sosial dan kemanusiaan, khususnya dalam penanganan bencana.

3. Negara

   Penelitian ini dapat menjadi dasar pengembangan sistem pemantauan bencana alam yang memanfaatkan informasi dari media sosial secara *real time*. Melalui pemanfaatan teknologi klasifikasi multilabel, informasi terkait bencana seperti lokasi kejadian, jenis kerusakan, jumlah korban, kebutuhan bantuan, hingga dampak sosial lainnya dapat diidentifikasi dan dikelompokkan secara otomatis. Dengan demikian, sistem ini memiliki potensi untuk menjadi alat pendukung yang efektif bagi instansi-instansi terkait seperti BNPB, BMKG, dan BPBD dalam memantau kondisi di lapangan dan merespons dengan cepat. Lebih jauh, sistem ini juga berpotensi mendukung pengembangan kebijakan kebencanaan berbasis data di masa mendatang, serta meningkatkan kolaborasi antara pemerintah, masyarakat, dan pihak-pihak lain dalam menghadapi dan menanggulangi bencana secara lebih adaptif dan responsif.

Halaman Ini Sengaja Dikosongkan